I. Response to Anonymous Referee #1

The authors would like to thank the Referee for the very helpful comments and suggestions. The comments have been taken into consideration in the revised manuscript. We answer all of them individually in the following.

1. General comments

1.1 Extrapolation to different climate conditions

The authors optimized parameters for single sites as well as for all sites of a PFT (multi-site). Then they applied these parameter sets for global model runs for global evaluation. Nevertheless, there is no analysis about how the optimized parameters can be used to predict NEE under different temporal, spatial or environmental conditions. A basic procedure in model optimization and evaluation is to optimize the model against one part of the data and evaluate it against the second part of the data. Specifically, I would like to see an analysis about how the multi-site parameter sets can predict NEE at sites that were not used during the optimization. Optimally, these sites should be also selected in a way that climate conditions are different from the optimization sites. This could provide us confidence how the model with the optimized parameter set performs under different climate conditions. This is an essential test for a model that is likely applied for climate change projections.

We acknowledge that "classical optimization procedures" keep one part of the dataset for validation. However, in our case, multi-site parameters transferability has not been evaluated at the site level given the small amount of sites for most PFTs (on the order of 10). It prevented to implement such an evaluation on a systematic and robust basis. The number of site is currently too small and one objective of the paper is to highlight the need to use all currently available sites. We acknowledge this limitation, and we thus improved the conclusion discussing the use of a larger set of site-years of flux data, in this text added to the conclusion (page 2984, line 8):

'*[...] would be beneficial. Using more site-years of flux data will also allow a systematic in-situ evaluation of the multi-site parameters across time periods, regions and climate regimes by separating training sites from evaluation sites. This procedure was not performed in this study due to the small number of sites for some PFTs, but remains essential to test a LSM used for climate projections. More generally, we suggest that the assimilation of FluxNet data […]*'

Additionally, the evaluations at the global scale against independent data give some insight on the ability of the multi-site parameters sets to improve the accuracy of the model at a much larger scales, likely implying different climate conditions. The time periods considered, 2000-2010 for NDVI and 1989-2009 for atmospheric CO2 concentrations, go beyond that covered by the flux data used (from 1996 to 2006) and therefore allow evaluating the parameters sets against periods of time the model was untrained for.

1.2 Evaluation of inter-annual variability

The authors state an improved model performance regarding inter-annual variability (IAV). I would be very interested in these results but unfortunately I cannot find any corresponding figures or tables. Could you please provide figures that demonstrate the improved IAV of ORCHIDEE regarding the following points?
- CO2: Demonstrate the improved IAV of CO2.
- NEE: Is there an improved IAV in comparison to sites with long time series?
- NDVI: Do you see improved IAV in mean growing season or peak NDVI? How do simulated

NDVI trends compare with observed NDVI trends?

The stated improvement of IAV in the abstract and conclusion only refer to the global simulation of atmospheric CO2 concentrations (p. 2962 lines 20-24 and p. 2984 lines 4-5). There is indeed no associated figure or table in this case, the results summarized over the CO2 records stations are described and discussed within the text p. 2982 lines 17-26. While no other claim is made in this study regarding interannual performances at the site level, or for the global simulations of FAPAR, we agree with the Referee that investigating the IAV of the NEE and the NDVI is also a crucial point. However, for the NDVI, computing the growing season mean or peak requires careful processing (especially over tropical regions with no clear seasonal cycle) that would have led to substantial new analysis. The same is true also for NEE. We thus decided to focus on the mean seasonal cycle, as a thorough analysis of the IAV is beyond the scope of this paper. It will be considered in a following study.

1.3 Comparison of modelled FAPAR with NDVI

Could you please provide some more details regarding the comparison of simulated FAPAR with NDVI? NDVI is also affected from non-vegetation changes like soil and snow reflectance. Especially, snow melt in spring can results in a fast increase in NDVI. In the computed FAPAR there is no snow effect and also no factor that accounts for background reflectance. Thus, the computed correlation is meaningless if one compares modelled FAPAR with NDVI that is affected by such non-vegetation related seasonal transitions. You should exclude NDVI observations that are possibly affected from snow or that are at the beginning or end of the growing season to draw more pure conclusions about model performance. Additionally, as the title states "to global simulations", I'm expecting to see some global model results and evaluations. Especially the NDVI comparison is highly aggregated into one table that does not provide much insight into model performance. I would rather expect maps and the corresponding discussion of correlation coefficients between modelled FAPAR and observed NDVI (weekly data, mean seasonal cycle, mean growing season comparison, trends). In which regions does the model perform well or why not?

We agree with the Referee that the comparison between model fAPAR and NDVI could be affected by several biases. In this analysis, observations contaminated by snow have been removed from the calculation of the correlation factors, using MODIS MOD13's quality filter. Following the Referee's suggestion, we minimized the effect of soil reflectance by applying a threshold criterion of 0.2 on NDVI, with little effect on the results except grasslands where the correlation coefficient became slightly lower, both with prior and multi-site parameterization. In the revised manuscript, the following sentence has been added page 2971, line 12:

'*Observations contaminated with snow cover were removed from the analysis, and we discarded NDVI observations below 0.2 in order to minimize the impact of bare soil reflectance.*'

In addition, Table 3 has been replaced by a boxplot figure were correlation factors are grouped by dominant PFT, with the following caption:

'*Figure 7. Correlation factor between weekly time series of modeled FAPAR and independent measurements of NDVI, for the 2000-2010 period. The results are grouped using the dominant PFT at each pixel, for global simulations with prior (green) and multi-site parameterization (blue). The central horizontal bar indicates the median value, the top and bottom of the boxes correspond to the first and last quartile, and the 5- and 95-percentile are given by the 'error bars'.*'

The description and discussion of the results in section 3.5.2 has been modified as follows:

'Figure 7 reports for each optimized PFT the correlation factors between weekly values of measured NDVI and modeled FAPAR during the period 2000-2010 (see sect. 2.4), for both the prior and optimized model. There is no result for BorDBF whose vegetation fraction never exceeds 40% in our case. All remaining six PFTs exhibit a higher median correlation factor when using the multi-site parameterization, which means that the modeled leaf seasonal cycle better matches the global scale observations. This median improvement seems to accurately reflect the overall trend for TempDBF-, BorENF- and C3grass-dominated pixels, while a larger inter-pixel variability is introduced in the case of temperate evergreen forests. The improved modeled seasonality is related to the more accurately simulated GPP at FluxNet sites after multi-site optimization, the latter being in turn partly driven by the improvement of the seasonal variations of simulated LAI. The dominant feature seems to be a shorter growing season length for TempDBF, which is consistent with the site-level simulations of GPP seasonality for this PFT (Fig. 5), and an earlier beginning of the growing season for C3 grasses (not shown). Note that this improvement also explains most of the change in the correlation factors in temperate and boreal evergreen forests, since these PFTs do not present a climate-driven leaf phenology in the current formulation of the ORCHIDEE model. Consequently, deciduous and herbaceous PFTs are the only significant contributors to the seasonal cycle at such a coarse resolution, even when these ecosystems are secondary and/or the understory within an evergreen-dominated forest. Lastly, the score for TropEBF remains poor because the model wrongly simulates the leaf renewal and the hydric stress during the dry season, as discussed in Sects. 3.1 and 3.4.'

1.4 Global total carbon stocks and fluxes

In optimization experiments, a parameter was introduced that regulates the initial soil and vegetation carbon pools in order to match the observations. I did not understand how this information was translated into the global model simulations. Did you account for spatially varying initial carbon pools? If yes, how? If not, how were carbon pools initialized and how might this affect model evaluation results? Additionally, I would like to see a table and discussion of global total carbon stocks and fluxes from the prior, single-site and multi-site experiments in comparison with estimated ranges from independent datasets.

This parameter is taken as site-specific, for this reason we did not extrapolate its optimized values and thus did not use the information for global simulations (page 2966 lines 23-29, page 2967 lines 1-3). The default initial carbon pool content is used in the latter case, obtained after a global spinup procedure similar to that done at the site level. It allows accounting for spatially-varying carbon pools since each pixel is spun up independently based on its vegetation, soil type and climate. The absolute carbon stocks could however be erroneous and would mostly affect the simulation of ecosystem respiration, and hence the modeled net carbon balance. For this reason, we did not analyze the global net carbon fluxes, nor the atmospheric CO2 concentration trend, but focused on the seasonal cycle and interannual variability of the latter. Global-scale comparisons or optimizations of ORCHIDEE parameter with spatialized carbon pools estimates/measurements and its use for long-term predictions are beyond the scope of this study. This is, however, a topic of active research in the context of building Carbon Cycle Data Assimilation Systems (Peylin et al., 2014, in preparation), where soil carbon pools might be optimized/scaled for an ensemble of "large-scale" regions.

1.5 Parameter variability and distributions

The manuscript misses a discussion on parameter uncertainty and variability. What is the spatial and within PFT-variability of parameters? How does a multi-site parameter value compare with the single-site variability? Which parameters were well constrained? Which are

uncertain? Are posterior parameter values plausible? I'm surprised not to see such results in a model-data fusion manuscript.

We agree that analyzing the values taken by the parameters would have been valuable. It has been done in detail in our previous multi-site study focusing on temperate deciduous broadleaf forests (Kuppel et al., 2012) and for the present PFTs in S. Kuppel's doctoral dissertation (Kuppel, 2012). However, we chose in this study to focus on analyzing the model outputs and model-data mismatches across sites for each considered PFTs, and the evaluation of the multi-site sets of parameters with global simulations. In this context, an in-depth analysis of parameter variability as performed, for example, in above references, would have made the manuscript very lengthy given the number of PFTs.

Besides, as stated in the manuscript page 2968 lines 15-17 we carefully prescribed the allowed parameters ranges in order to keep physically-sound posterior parameter values, at the risk of reducing the leverage the optimization on the modeled fluxes.

Lastly, the grouped parameter uncertainty is analyzed in section 3.3, from an output perspective. This analysis does not allow to directly identify which parameters/processes are better constrained, but it does provide insight into the weight of parameter uncertainty as a whole in the separate uncertainty budget of NEE and LE simulation within and between PFTs, and its implications regarding the limits of the current model structure.

1.6 General discussion and significance of the study

The discussion of model limitations is currently distributed over the entire results section. I would suggest adding another section before the conclusions that summarizes the limitations and potential need for improvement of the model that were identified in optimization experiments. Additionally, this section should also discuss the relevance of this work for other modelling groups or for model-data fusion in general. This can potentially improve the importance and impact of this manuscript for other groups.

We agree and thank the Referee for this suggestion. In the revised manuscript, a section 3.6 entitled '*Limitations of the current approach: summary and discussion*' has been added (see below). Note that the emphasis is put on discussing the limitations of our method, as we feel that the relevance of this work was already stated in our conclusions.

'*The limitations to our model-data fusion method highlighted throughout the results section are of three kinds, somewhat interlocked: 1) within the limits of the model structure, 2) how adequate the chosen set of optimized parameters was and 3) how close to the optimal values the optimization algorithm tuned these parameters.*

Taking these items in reverse order, we first acknowledge that using a variational optimization algorithm with a model with non-linearities might expose to miss the global minimum of the cost function, and indeed a few obvious convergence failures cases have been found for some single-site optimizations in TropEBF, TempENF, and boreal forests. Some functions of the ORCHIDEE model could potentially be linearized to generate a more accurate tangent linear model –and to advantageously avoid to use finite-differences for some phenological parameters (see Sect. 2.1)–, while remaining coherent with the model's philosophy. It might imply a demanding effort of model recoding, but it has already been done for another LSM (Knorr et al., 2010). Alternatively, stochastic optimization approaches could yield better convergence, as they can circumvent the linearity constraint. While a single-site model-data fusion study with the same LSM showed advantageous results for a genetic Monte-Carlo-based technique over its variational counterpart (Santaren et al., 2013), no major difference was found by (Ziehn et al., 2012) between

Monte-Carlo and gradient-based approaches when optimizing a simpler LSM with atmospheric CO_2 observations. In the case of a multi-site optimization efforts, we suggest that the cost function "smoothing" discussed in Sect. 3.2 could make the convergence efficiency less sensitive to the choice of the minimization approach, while keeping in mind the much lower computational time required in the variational case.

Second, the number of optimized parameters remains somewhat modest as compared to the diversity of processes modeled in the ORCHIDEE model. Our choice was partly driven by a model sensitivity criterion, while the actual leverage of an optimized parameter on model outputs also depends on the uncertainty associated to this very parameter (Dietze et al., 2014). It can result in selecting some parameters that are already reasonably well known but that have medium-to-high model sensitivity and thus with low overall leverage, while poorly known parameters with mild-to-low model sensitivity could have a comparatively higher value for the optimization. In addition, as our focus was on the carbon cycle, only a few water-and-energy-related parameters were considered. Notably, the correction of LE partly benefited from that of NEE via transpiration, but the soil evaporation optimization was neglected despite being a significant -and debated- player of the terrestrial water cycle (Schlesinger and Jasechko, 2014).

The third hindering factor to simulating carbon and water fluxes close to their true value is the "observation error", i.e. the uncertainty arising from the simplification needed to make ecosystem functioning fit within explicit equations plus the error made associated to the measurements, fluxes and meteorological forcing included. Although this error is rarely quantified in model-data fusion efforts, model-data fit analyses and uncertainty budgets showed in this study that the relative importance of this observation error greatly varies from one PFT to another –and is potentially dominated by the model error component in the case simulations at flux towers sites (Kuppel et al., 2013). It is the highest in tropical evergreen broadleaf forests, where parameter optimization will likely be of limited help until a more realistic phenological scheme is implemented. Regarding the simulations of LE in general, the small amount of related parameters optimized makes it difficult to assess to which extent the nearly-unchanged flux uncertainty comes from the parameter scarcity or structural inaccuracies in the model, stressing again the need for a better consideration of water and energy cycles together with that of carbon in future model-data fusion efforts.'

2. Specific comments:

page 2962, line 3-4: Please write "net ecosystem exchange" to introduce the abbreviation NEE.

Corrected.

page 2966, line 27-29: I don't understand why the multiplier for LAI was not applied for deciduous PFTs. It should be the same like for evergreen and herbaceous PFTs that the maximum annual coverage of deciduous PFTs depends on the site history. Can you please clarify this?

The parameter LAI_{init} is a multiplier only for the initial LAI value, scaling the initial foliar cover from the output of the spinup (in a analogous fashion as K_{soilC} for the initial soil carbon stock). While the value of LAI at the first time step of the simulation significantly determines the later foliar covers of evergreen and herbaceous PFTs, deciduous PFTs have their LAI almost entirely reinitialized each year when leafs fall. In our view, the initial LAI of deciduous PFTs has thus little leverage on their foliar cover variation, all the more that it is very low in early January when simulations start, which is why we did not optimize LAI_{init} in these cases.

page 2968, line 15-17: I don't understand this sentence. Is this reproducible?

The allowed range of variation for each parameter derives from ecophysiological considerations when the parameter possesses a direct physical definition (e.g., the maximum carboxylation rate), while for purely empirical parameter the focus has been set on maintaining the response functions associated to it within reasonable boundaries. While plant trait databases were used as much as possible for physically-meaningful parameters, personal communications with ORCHIDEE modelers has been essential to adapt literature knowledge to ORCHIDEE specificities. In order to make the experiment reproducible, a new supplementary Table S1 gives all the parameter intervals used in the study.

page 2969, line 7-10: How was the optimization done, if the remaining 30% of the grid cell were covered by another PFT (i.e. understory, grass?). Was the minor PFT represented in the optimization? If not, what is the risk that the dominant PFT accounts for changes that are due to the minor PFT? Or were both PFTs optimized at the same time or sequentially?

Only the dominant PFT was considered for optimization, and indeed part of the optimization results could be artifacts 'compensating' for an inadequate modeling of the minor PFT. However, preliminary tests optimizing minor PFT(s) as well were performed on a subset of sites (at least one for each dominant PFT) in 'single-site' mode and showed no significant differences in the model-data fit improvements. Second, the aim of this study is to assess the robustness of PFT-generic approach against a site-specific one within the parameter optimization framework, i.e. how reasonably realistic it is to consider sites with a same dominant PFT as clustered information, to ultimately correct the ORCHIDEE model for global-scale simulations. This goal notably implies having a single parameter set for a given PFT. The C3 grasslands are a minor PFT at many sites considered in this study, but is also optimized as a dominant PFT -and at most of the corresponding sites the coverage is 100%, thus yielding several parameters sets for this PFT if all represented PFTs were to be included in the multi-site optimization. Also, the understory or minor PFT(s) are not always the same between sites of a same dominant PFT, here again reducing the genericity of a multi-site optimization considering all represented PFTs. In our view, these reasons justify neglecting the minor PFTs in the optimization as a reasonable approximation.

page 2971, line 19-21: How were snow or albedo changes considered?

Please refer to our response to general comment 1.3.

page 2971, line 25: Does this refer to the coverage of the dominant PFT or if total coverage of all PFTs? Did you evaluate also in grid cells that had a mixture of several PFTs? If not, why not? If yes, how was the model performance?

It refers to the coverage of the dominant PFT (bare soil included), the sentence has been modified and now reads: '*[...] we restrict our correlation computation to the model boxes where the dominant PFT's cover fraction exceeds 50 % and where [...]*'. In the analysis, the focus is made on the dominant PFT, regardless if the latter has a fraction equal to 100% or if it shares the pixel with 'secondary' PFTs. We acknowledge some limits to this approach as for example improvements of the simulated phenology in boreal evergreen forests might actually reflect improvements of the secondary deciduous cover and/or the herbaceous under story (page 2983, lines 13-18).

page 2973, line 12-13: Why there were only small improvements in evergreen PFTs? Could this be linked to the phenology routine?

The inner structure of the phenology routine in ORCHIDEE is indeed a potentially important factor explaining the poor results in tropical evergreen broadleaf forests, as discussed page 2975 lines 13-29, page 2976 lines 1-15, page 2978 lines 15-2, page 2979 lines 18-21, and page 2983 lines

18-20. Regarding boreal evergreen needleleaf forests, the reduction of model-data RMSD was indeed among the lowest for NEE, noting however that the average prior misfit was also the lowest among all the PFTs (even after optimization of these other PFTs, Fig. 1a), as was the prior model-data bias (Fig. 1b). In both cases, the limited leverage of parameter optimization might indicate that further improvements requires reconsidering the structural equations of the ORCHIDEE model, among them those linked to phenology.

page 2982, line 17: Please demonstrate this with a corresponding figure or table.

In order not to overload the manuscript with figures, we chose not to show any figures for the 53 records at CO2 stations except for Fig. 6 to discussing a few relevant sites, so that the related results are directly summarized in the main text. This is even truer now that Figure 7 was added, following the Referee's suggestion (see response to comment 1.3).

page 2983, line 15-16: Why? Is this because evergreen PFTs don't have a phenology in your model and there are no seasonal effects of snow cover?

Indeed, no seasonal effects of snow cover are considered, as it will be stated in the revised manuscript (see response to general comment 1.3). Secondly, the Referee is again right in pointing out the lack of a phenology model for evergreen PFTs. Only winter dormancy is considered, as photosynthesis is not permitted if the monthly temperature remains below a PFT-dependent threshold (Krinner et al., 2005), and the leaf turnover throughout the year. The sentence has been modified in the revised manuscript, and now reads:

'Note that this improvement also explains most of the increased correlation factors in temperate and boreal evergreen forests, since these PFTs do not present a climate-driven leaf phenology in the current formulation of the ORCHIDEE model. Consequently, deciduous and herbaceous PFTs are the only significant contributors to the seasonal cycle at such a coarse resolution, even when these ecosystems are secondary and/or the understory within an evergreen-dominated forest.'

Table 1: There are no values underlined but some are in bold font. Please clarify.

The underlining was replaced by bold font during the manuscript's typesetting, but the caption of Table 1 had not been updated accordingly. We have modified it in the revised version of the manuscript: '*Parameters of ORCHIDEE optimized in this study. The prior values are given for each PFT, and multi-site posterior values are in bold font. A hyphen means that the parameter is not optimized, spinup that the spinup value is taken, and* **site** *that the posterior value is site-specific.*'

Table 2: This table is very long but not very informative. I would suggest moving this to the appendix or supplementary material like the table for the CO2 stations.

The table has been moved to the supplementary material as Table S2 with the associated references, and Table S1 has become Table S3 (a new Table S1 with parameter intervals as been added as well, see above).

Table 3: This table is not very informative. The differences are small. Could you please provide an estimate of the significance of these differences? Even better would be a map of correlations or boxplots of the global distributions of correlations.

Please see response to comment 1.3

Figures 1 and 2: It is not clear if (a) and (b) refer to the mean seasonal cycle or to the full

length of the time series. Please add a legend with colours and symbols to the plot to improve the readability of the figure. I would not expect biases in the posterior of single site optimizations. What are the reasons for these biases? The y-axis scale in Fig. 1 c for TempDBF is not very different; thus please use the same scale in order not to confuse the reader.

In both Fig. 1 and 2, (A) and (B) refer to the full length of the time series, and we completed the captions. Besides, color legends have been added to both figures. We finally considered that using different symbols for each PFT was somewhat unnecessary, as the results are horizontally grouped by PFT, all the more that a symbol legend would overload the graphics. For these reasons, all the symbols have been changed to open circles in Figs. 1 and 2, as well as in Fig. 4, The respective captions have been modified accordingly:

'*Figure 1. Model-data **(A)** RMSD and **(B)** bias for the daily NEE time series at each site (filled circles), grouped and averaged by PFT (horizontal bars), in three cases: prior model (green), multi-site optimization (blue) and single-site optimization (orange). **(C)** PFT-averaged mean seasonal cycle of NEE, for the training observations (black) and the three aforementioned cases, smoothed with a 15-day-moving-average window.*

*Figure 2. Model-data **(A)** RMSD and **(B)** bias for the daily LE time series at each site (filled circles), grouped and averaged by PFT (horizontal bars), in three cases: prior model (green), multi-site optimization (blue) and single-site optimization (orange). **(C)** PFT-averaged mean seasonal cycle of LE, for the training observations (black) and the three aforementioned cases, smoothed with a 15-day-moving-average window.*

*Figure 4. Uncertainty of simulated daily **(A)** NEE and **(B)** LE fluxes. For each PFT, the horizontal lines give the average of the individual site values (filled circles), in three cases: prior model (green), multi-site optimization (blue) and single-site optimization (orange).*'

In the revised version of the manuscript, the y-scale of NEE in TempDBF (Fig 1C), has been evened with that of the other PFTs. The same has been done in Fig. 5A regarding the y-scale of GPP.

Finally, a posterior model-data bias is not surprising in our opinion, even for single-site optimizations, primarily because model-data bias is not the metrics minimized by the inversion algorithm, the latter being RMSD (page 2975, lines 1-5). Hence we are not sure why the Referee would not expect any posterior biases.

Figure 3, 4, 5: Please add colour legends to all figures.

In the revised manuscript, color legends have been added to Figs. 1, 2, (see above comment), as well as 3, 4, and 5.

Figure 5 and 6: It would be valuable information to have some model performance measures (RMSD, correlation) included above the cycles for each PFT.

We assume the Referee is referring to Figs. 1 and 2. In our opinion, the discussion on how well the mean seasonal cycles of NEE and LE simulated by ORCHIDEE compares to the observations, on PFT average, is quantified in Fig. 3 and the associated analysis (page 2974 lines 8-28, page 2975, and page 2976 lines 1-15). Note that the phase coefficient (page 2970, Eq. 3) was preferred to the correlation in non-tropical PFTs, since the focus was put on evaluating the accuracy of the simulated limits of the growing season. For this reason, we do not think necessary to add further performances metrics to the seasonal cycle of Figs. 1 and 2, considering in addition a potential overload of these figures.

References

Dietze, M. C., Serbin, S. P., Davidson, C., Desai, A. R., Feng, X., Kelly, R., Kooper, R., LeBauer, D., Mantooth, J., McHenry, K. and others: A quantitative assessment of a terrestrial biosphere model's data needs across North American biomes, J. Geophys. Res. Biogeosciences, 119(3), 286–300, 2014.

Knorr, W., Kaminski, T., Scholze, M., Gobron, N., Pinty, B., Giering, R. and Mathieu, P.-P.: Carbon cycle data assimilation with a generic phenology model, J. Geophys. Res. Biogeosciences 2005–2012, 115(G4), 2010.

Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogée, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S. and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, Glob. Biogeochem. Cycles, 19(1), 2005.

Kuppel, S.: Assimilation de mesures de flux turbulents d'eau et de carbone dans un modèle de la biosphère continentale, PhD thesis, Université de Versailles Saint-Quentin-en-Yvelines., 2012.

Kuppel, S., Chevallier, F. and Peylin, P.: Quantifying the model structural error in carbon cycle data assimilation systems, Geosci. Model Dev., 6(1), 45–55, 2013.

Kuppel, S., Peylin, P., Chevallier, F., Bacour, C., Maignan, F. and Richardson, A. D.: Constraining a global ecosystem model with multi-site eddy-covariance data, Biogeosciences, 9(10), 3757–3776, 2012.

Peylin, P., Bacour, C., MacBean, N., Leonard, S., Maignan, F., Thum, T., Chevallier, F., Ciais, P., Cadule, P. and Santaren, D.: How best to optimize a global process-based carbon land surface model?, in EGU General Assembly Conference Abstracts, vol. 16, p. 10302, 2014.

Santaren, D., Peylin, P., Bacour, C., Ciais, P. and Longdoz, B.: Ecosystem model optimization using in-situ flux observations: benefit of monte-carlo vs. variational schemes and analyses of the year-to-year model performances, Biogeosciences Discuss., 10(11), 18009–18064, 2013.

Schlesinger, W. H. and Jasechko, S.: Transpiration in the global water cycle, Agric. For. Meteorol., 189, 115–117, 2014.

Ziehn, T., Scholze, M. and Knorr, W.: On the capability of Monte Carlo and adjoint inversion techniques to derive posterior parameter uncertainties in terrestrial ecosystem models, Glob. Biogeochem. Cycles, 26(3), 2012.

II. Response to David Schimel

The authors would like to thank D. Schimel for his comments and the profound perspective suggested. It has been taken into consideration in the revised manuscript.

This is a very nice contribution and a significant advance in the practice of ecosystem carbon data assimilation. The work is well done, conforms to the state of practice, advances the field and is clearly presented for the most part.

I have one perspective to add. The approach taken is directly analogous to similar estimation approaches in meteorology, and is useful but in a sense not informative. Consider the actual situation being modeled. Ecosystems, far from being a continuous field of "green slime" are in

fact made of up of billions of individual plants, and even more bazillions of leaves and microbes etc. Within a single model plant functional type that can be up to tens of thousands of species, each with slightly or significantly different parameter values for the model equations. The goal of assigning PFTs and biomes is to reduce the unmanageable dimensionality of this variation to a reasonable degree, and the study presented here shows that using replication of flux sites –even though they do not systematically or randomly sample this variability– helps improve overall model performance.

However, this analysis does not take into account any covariance structure associated with the underlying structure of parameter variation associated with species or functional variation. Treating parameter variation as a random field is a reasonable first assumption but is almost certainly not true. It would be interesting to consider or speculate on how such an analysis would be done if more detailed information on plant parameter distributions were available to weight extrapolation from a limited set of towers. In any case, adding a description of the conceptual situation in which this assimilation is taking place would be useful. As ecosystem data assimilation progresses, making a transition to a more biologically sophisticated underlying model will be critical.

This comment is indeed very accurate. As the long-term aim of model-data fusion is here to assess the structural limits of ecosystem models based on PFT concept and to guide later conceptual developments, ignoring parameter covariances will ultimately bias the results and could for example wrongly attribute model deficiencies. In the revised manuscript, the conclusions have been completed to include this idea, as follows:

'*[…] maximize this improvement. In parallel, by using a diagonal prior covariance matrix for parameter error, within a same PFT and across PFTs, we implicitly assumed that all parameters could in principle be efficiently corrected as independent random distributions. It ignores the fact that a covariance structure interlinking the optimized parameterization would be necessary to translate the interconnectedness of ecophysiological processes within a given PFT. For instance, the allocation of carbon within the plant reservoirs depends on specific allometric relations and on photosynthesis rate; these relations would need to be embedded in the prior parameter error covariance matrix. Additionally, the influence of nearby individuals of other PFTs (e.g., the understory) should be accounted for when correcting parameters of a given PFT. Together with a simultaneous optimization of several PFTs, building standard spatialized parameter covariance tables from databases of plant traits and soil characteristics (e.g., (Kattge et al., 2011)) and 'preliminary' posterior multi-site parameter error covariance matrices (e.g., supplementary material of (Kuppel et al., 2012)) might soon become necessary to consistently apply model-data fusion to more sophisticated mechanistic ecosystem models.* '

References

Kattge, J., Diaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J. and others: TRY–a global database of plant traits, Glob. Change Biol., 17(9), 2905–2935, 2011.

Kuppel, S., Peylin, P., Chevallier, F., Bacour, C., Maignan, F. and Richardson, A. D.: Constraining a global ecosystem model with multi-site eddy-covariance data, Biogeosciences, 9(10), 3757–3776, 2012.

III. Response to Matthew Smith

The authors would like to thank M. Smith for his helpful comments and suggestions. They have been taken into consideration in the revised manuscript. We answer all of them individually in the following, merging the two parts of the posted review.

General comments

This study constrains a global ecosystem model (estimates of the most likely parameters) using multiple datasets from multiple sites and shows resulting improvements in model predictive performance in predicting the CO2 fluxes as well as other performance metrics at multiple sites. Investigations of where predictive performance has been improved or made worse reveal insights into how the process has influenced the general applicability of the model – it has improved at capturing CO2 fluxed at tropical and temperate sites but has identifiable weaknesses in predicting tropical evergreen broadleaf forest dynamics which leads to the identification of new areas for research. They also illustrate the efficacy of the model at predicting CO2 flux dynamics for a wider set of test sites and conduct a global scale evaluation. In sum, this to me is an excellent end-to-end analysis of the costs and benefits of undertaking this more sophisticated and improved model fitting approach and I recommend it for publication.

Specific comments

It is perhaps worth noting in the results and discussion that, as far as I can see, none of your effects from parameter estimation lead to qualitative differences in the predictions of the model. They simply lead to quantitative improvements. This implies to me that when we are moving towards a situation that we have multiple data-constrained DGVMs being used in climate simulations, each will demonstrably predict the present day data better, but their predictions of the future, and the differences in their predictions of the present, will still vary widely. This to me implies that while you are improving the parameterisation under the assumed model structure, you are not improving the assumed model structure to make it better suited to modelling reality and it is this which needs more focus of the attention of DGVMers.

Although the improvements are indeed mostly quantitative, note for example that the interannuality of the simulated atmospheric CO2 concentrations has been improved, although modestly. This results emerges for time series much longer (20 years) than those used for the ORCHIDEE LSM optimization. It suggests that although parameter optimization always remains by definition within the limits of the model structure, simulations outside the time periods used for optimization can be corrected with this tool, to some extent. Further work is needed to assess more accurately how large exactly the aforementioned extent is, in the case of the ORCHIDEE model see Santaren et al., (2013)**.**
However, we agree with the reviewer that a crucial question is whether applying data assimilation to all models (used for instance in the CMIP5 exercise for the IPCC report) would decrease or not the spread in the future predictions of the carbon cycle and, consequently, in climate predictions. Although the differences after optimization of the ORCHIDEE model lead to quantitative improvement but no large qualitative changes, it is difficult to assess their impact under climate change. For instance, an ongoing study based on the assimilation of flux tower data (this work) and satellite NDVI data, with the same model, led to significant changes of the soil carbon stocks after 2050 when used with future climate projections (from CMIP5). The changes appeared when climate warming reached a certain level, where the modified parameters start to induce large flux

differences (heterotrophic respiration). The non-linearity of the model is in this case crucial.

1. Table 1 legend - nothing is underlined, I think you mean bold

Yes, this typesetting mistake has been corrected in the table caption for the revised version:
'*Table 1. Parameters of ORCHIDEE optimized in this study. The prior values are given for each PFT, and multi-site posterior values are in bold font. A hyphen means that the parameter is not optimized, spinup that the spinup value is taken, and* **site** *that the posterior value is site-specific.*'

2. Could you please indicate for your figures and tables (e.g. Table 3) whether these assessments are for independent evaluation data or for the data the model was trained to.

The legends of Figs. 1 and 2 have been modified to emphasize that we present the training data as observations, while Figs. 5, 6 and 7 (the latter replacing Table 3 in the revised manuscript) have been modified state that they present evaluative (and independent for Figs. 5 and 6) data:

'*Figure 1. Model-data* ***(A)*** *RMSD and* ***(B)*** *bias for the daily NEE time series at each site (filled circles), grouped and averaged by PFT (horizontal bars), in three cases: prior model (green), multi-site optimization (blue) and single-site optimization (orange).* ***(C)*** *PFT-averaged mean seasonal cycle of NEE, for the training observations (black) and the three aforementioned cases, smoothed with a 15-day-moving-average window.*

Figure 2. Model-data ***(A)*** *RMSD and* ***(B)*** *bias for the daily LE time series at each site (filled circles), grouped and averaged by PFT (horizontal bars), in three cases: prior model (green), multi-site optimization (blue) and single-site optimization (orange).* ***(C)*** *PFT-averaged mean seasonal cycle of LE, for the training observations (black) and the three aforementioned cases, smoothed with a 15-day-moving-average window.*

Figure 5. PFT-averaged mean seasonal cycles of ***(A)*** *the photosynthetic carbon flux and* ***(B)*** *the respiration flux, smoothed with a 15-day-moving-average window. The simulations using prior (green), single-site (orange) and multi-site (blue) parameterizations are compared to the evaluative observation-derived flux estimates (black).*

Figure 6. Detrended mean seasonal cycle of the atmospheric CO_2 concentrations at ***(A)*** *Alert,* ***(B)*** *South Pole and* ***(C)*** *Mauna Loa locations during the 1989-2009 period: the optimization-independent concentrations records (black) are compared to simulations where the biospheric contribution is calculated using the ORCHIDEE model with default (green) and multi-site (blue) parameterization, with the model-data RMSD given between brackets.* ***(D)*** *Regional contributions to the mean seasonal cycle simulated at Alert.*

Figure 7. Correlation factor between weekly time series of modeled FAPAR and independent measurements of NDVI, for the 2000-2010 period. The results are grouped using the dominant PFT at each pixel, for global simulations with default (green) and multi-site parameterization (blue). The central horizontal bar indicates the median value, the top and bottom of the boxes correspond to the first and last quartile, and the 5- and 95-percentile are given by the 'error bars'.'

3. ALL FIGs - it is not clear to me whether the figures relate to an average across PFTs, which specific years were considered or anything. While these fits look good, I have little idea what places in space and time they specifically relate to. You need to improve the legends to these figures to explain this.

The distinction between site-level and PFT-averaged values are made in the caption of Figs. 1, 2, 3, 4 and 5. Color legends have been added to Figs. 1, 2, 3, 4, and 5 to better distinguish whether the displayed quantities relate to prior, multi-site or single-site cases, or to the observations. Finally, to

make clearer whether the full-length of the time series or the mean seasonal cycle of carbon and water fluxes are considered, the captions of Figs. 1, 2 and 3 have been modified in the revised manuscript:

Figs. 1 & 2: see above response.

'*Figure 3. PFT-averaged model phase coefficient versus model-to-data amplitude ratio, for the detrended smooth seasonal cycles of* ***(A)*** *NEE and* ***(B)*** *LE fluxes. Simulations using prior parameters (green) are compared to multi-site (blue) and single-site (orange) optimizations, with the measured reference indicated by the intersection of the dashed lines.*'

References

Santaren, D., Peylin, P., Bacour, C., Ciais, P. and Longdoz, B.: Ecosystem model optimization using in-situ flux observations: benefit of monte-carlo vs. variational schemes and analyses of the year-to-year model performances, Biogeosciences Discuss., 10(11), 18009–18064, 2013.